Decomposition of terminology graphs for domain knowledge acquisition.

Fidelia Ibekwe-Sanjuan, Eric Sanjuan, Michael Vogeley

▶ **To cite this version:**

Fidelia Ibekwe-Sanjuan, Eric Sanjuan, Michael Vogeley. Decomposition of terminology graphs for domain knowledge acquisition.. 17th ACM conference on Information and knowledge management (CIKM '08), Oct 2008, Napa Valley, California, United States. pp.1463-1464, 10.1145/1458082.1458334 . hal-00636039

HAL Id: hal-00636039
https://hal.science/hal-00636039

Submitted on 26 Oct 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Decomposition of Terminology Graphs for Domain Knowledge Acquisition

Fidelia Ibekwe-SanJuan
ELICO - University of Lyon 3
France
ibekwe@univ-lyon3.fr

Eric SanJuan
LIA - University of Avignon
France
eric.sanjuan@univ-avignon.fr

Michael S. Vogeley
Department of Physics - Drexel University
Philadelphia, PA 19104
USA
msv23@drexel.edu

ABSTRACT

We propose a graph decomposition algorithm for analyzing the structure of complex graph networks. After multi-word term extraction, we apply techniques from text mining and visual analytics in a novel way by integrating symbolic and numeric information to build clusters of domain topics. Terms are clustered based on surface linguistic variations and clusters are inserted in an association network based on their intersection with documents. The graph is then decomposed based on atom graph structure into central (non-decomposable) atom and peripheral atoms. The whole process is applied to publications from the Sloan Digital Sky Survey (SDSS) project in the Astronomy field. The mapping obtained was evaluated by a domain expert and appeared to have captured interesting conceptual relations between different domain topics.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content analysis and indexing; H.3.3 [**Information Storage and Retrieval**]: Clustering

Keywords

Knowledge acquisition, Mapping knowledge domains, Graph decomposition, Information visualization, User evaluation

1. INTRODUCTION

We address the knowledge acquisition needs of domain specialists by automatically mapping the structure of a specialty in the field of Astronomy, that of the Sloan Digital Sky Survey (SDSS) project. We designed a text mining system called TermWatch which extracts terms (multi-word noun phrases) from running texts and clusters them based on an integrated approach combining symbolic (linguistic) and numerical relations (co-occurrence). It relies on a graph-based approach using a hierarchical clustering algorithm named CPCL (Classification by Preferential Clustered Link)[2]. The results are evaluated by a domain expert involved in the SDSS project who is also co-author of this paper.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIKM'08 October 26–30, 2008 Napa Valley, California USA
Copyright 2008 ACM 978-1-59593-991-3/08/10 ...$5.00.

2. METHODOLOGY

2.1 SDSS dataset

The corpus consisted of bibliographic records of peer-reviewed journal papers related to the SDSS project, published between 1991 and 2006. The records were retrieved from the Web of Science. 1,293 bibliographic records were thus extracted. While this collection does not constitute the whole set of publications on SDSS, it represents a body of peer reviewed publications indexed by the ISI.

2.2 Feature extraction and selection

TermWatch extracts multi-word terminological noun phrases based on our set of morpho-syntactic rules and POS (part-of-speech) information by (TreeTagger, Schmid 1994). We devised an adapted weighting function which takes the geometric mean (G_{mean}) of the inertia induced by two $tf.idf$ functions.

2.3 Term Clustering

The next step is to build a network of semantic variants based on linguistic relations which relate synonymous terms (orthographic variants, WordNet synonyms), siblings (expansions of the same terms) or create a loose type of association. In a first phase, tight semantic relations are used to form connected components of the input network. These components capture the synonymous variants of the same concept, thus ensuring that semantically equivalent terms or near equivalent are not dispersed in several clusters at the end of the process. The next phase is the clustering proper using a coefficient to compute strength of between components. The CPCL algorithm has been evaluated against variants of hierarchical clustering and k-means and was found to produce more homogeneous clusters. See [3] for more details.

2.4 Atom decomposition of association graphs

It involves two major steps: the construction of an association graph from any two information units and graph decomposition. To analyze assocation in documents between clusters, we compute overlapping atoms instead of disjoint connected components.

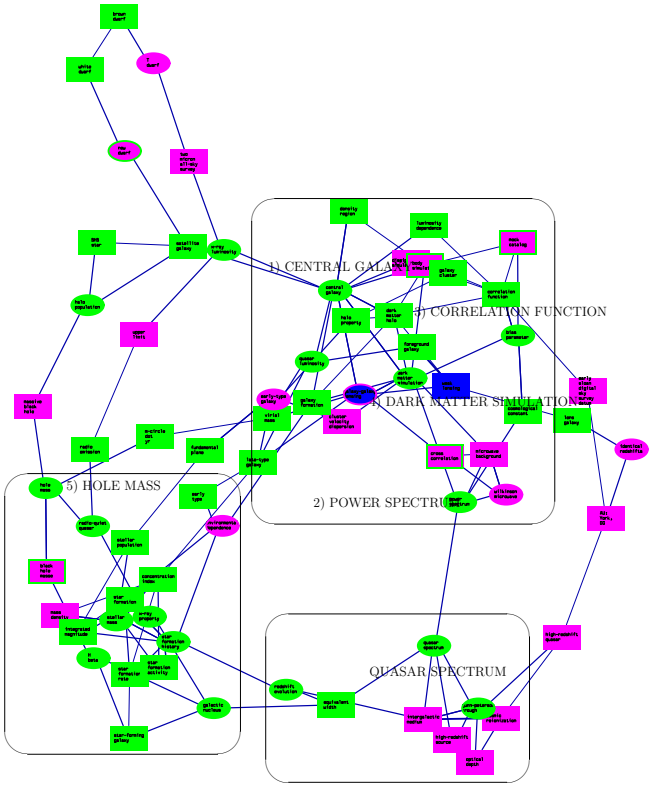

Figure 1: Central atom

To analyse such network, we rely on the intrinsic graph structure.

For that we use the the concept of graph atom that can be defined based on (a, b)-clique separators. These are complete subgraphs such that there exists two vertices a, b such that any path from a to b necessarily contains at least one element in the separator. We shall say that a graph is clique inseparable if there is no subgraph that is a complete separator. We shall call atom of a graph any connected maximal subgraph that is clique inseparable. By definition, an atom A of G_k contains at least one complete separator S of G_k. However S is not a separator of A. Atoms overlap if they contain the same separator of G_k. The decomposition of G_k in atoms is unique and it can be decomposed in $O(|V||E|)$. In previous experiments [1], we observed that graphs of the form $G_k(V, E)$ for k between 1 and 3 have a central atom with long cycles that involves almost 50% of the vertices and numerous peripheral atoms of small size that are almost chordal (circles have less than three elements).

3. RESULTS

3.1 The central atom in SDSS research

This graph portrays a subset of inseparable nodes as defined previously. In terms of domain knowledge, it can be seen as a set of tightly interwoven topics that could be core in the domain.

Figure 1 shows the general layout of the graph of associations between clusters before atom decomposition using the AiSee graph display package (http://www.aisee.com).

3.2 Evaluation

The clusters are labeled automatically by the system. The central atom graph was presented to the expert for evaluation. Globally, the structure of the central atom was found to be coherent, in some cases highlighting interesting and unexpected connections between domain topics.

The most prominent node labelled "central galaxy" is positioned at the center of the central atom. The fact that the system placed this cluster in a central position was judged an unexpected but interesting finding by the domain expert. It also made sense because new theoretical ideas about how galaxies populated dark matter halos (the "Halo Occcupation Distribution" model), observational studies to determine the masses of central galaxies by observing velocities of satellites, and related observational studies using weak lensing, all focus on "central galaxies". Thus, the system was successful in highlighting as such, a passageway in the domain without external domain knowledge.

"Power spectrum" the is 2nd node by betweenness centrality and is strategically placed between the first group on "central galaxy" and a group of topics in the southern region of the central atom. Expert evaluation confirmed that the links around power spectrum were valid.

"Correlation function" is ranked 3rd by betweenness centrality. This cluster shares an indirect link with "central galaxy" via "dark matter halo". Expert evaluation confirmed that this term is in some cases related to the power spectrum statistic (in such cases they are the Fourier transform of one another). "Dark matter simulation" is 4th by betweenness centrality. This cluster is linked to other clusters labeled "galaxy-galaxy lensing, galaxy formation, body simulation". They focus on recent techniques or functions to study galaxies and measure dark matter haloes.

"Hole mass" ranked 5th by betweenness is located at the extreme left of the graph and seems to be part of a network of publications on the study of black holes.

4. CONCLUSION

We designed a system for automatically mapping the structure of a specialty field. Our graph decomposition method has accurately isolated in a central atom the core topics and correctly related them without external semantic knowledge. The structure of the core topics could not have been grasped by reading up publications or by a pre-existing domain knowledge representation.

5. REFERENCES

[1] M. D. Biha, B. Kaba, M.-J. Meurs, and E. SanJuan. Graph decomposition approaches for terminology graphs. In *MICAI*, pages 883–893, 2007.

[2] F. Ibekwe-SanJuan. A linguistic and mathematical method for mapping thematic trends from texts. In *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI)*, pages 170–174, Brighton, UK, August 1998.

[3] E. SanJuan and F. Ibekwe-SanJuan. Text mining without document context. *Inf. Process. Manage.*, 42(6):1532–1552, 2006.